

Annotation Enrichment Analysis: An Alternative Method for Evaluating the Functional Properties of Gene Sets

Kimberly Glass^{1,2,3,*}, and Michelle Girvan^{2,3}

¹*Department of Biostatistics and Computational Biology,
Dana-Farber Cancer Institute, Boston, MA, USA*

²*Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA*

³*Department of Physics, University of Maryland, College Park, MD, USA*

⁴*Institute for Physical Science and Technology, University of Maryland, College Park, MD, USA*

Functional annotation databases provide a wealth of information for evaluating and interpreting biological systems. These databases generally exhibit non-uniform annotation properties in which many genes are annotated to a few non-specific biological functions, but only a few genes are associated with each of the numerous highly-specific biological functions. We investigate how these annotation properties influence the predictions made when using standard overlap statistics to determine the functional enrichment of gene sets. We find such approaches are strongly biased toward over-estimating the overlap significance between a gene signature and a functional category if either is associated with an unusually high number of annotations. More specifically, we determine the statistical significance of the overlap between randomly generated sets of genes and genes annotated to Gene Ontology categories using Fisher's Exact Test (FET), a statistical measure commonly employed in functional enrichment analysis, and show that the significance level of the association between a gene set and a functional category is positively correlated with the both the number of annotations made by genes in that gene set and the number of annotations to that functional category. Furthermore, we point out that many published gene signatures include a large number of highly annotated genes. To correct for this annotation number bias, we develop Annotation Enrichment Analysis (AEA) and use it to evaluate the functional significance of published gene signatures. We show that, especially when fully considering the structural properties of Gene Ontology annotations, AEA is able to predict biologically meaningful results, most of which are contained in FET, but are obscured by the many false-positive enrichment scores that occur in FET due to annotation number bias. We suggest that AEA should be used either alongside or in lieu of traditionally-used statistics when evaluating the functional enrichment of GO categories in gene sets so as to more accurately capture the biological functions represented by those gene sets.

1. INTRODUCTION

1.1. Motivation

Evaluating the functional properties of gene sets has been used both to verify that the genes implicated in a biological experiment are functionally relevant to the system in question [28] and to discover unexpected shared functions between those genes [31, 41]. This type of analysis has become a routine step in understanding high-throughput biological data [28, 53]. It is therefore important that functional enrichment analysis return results that are both statistically sound and biologically meaningful.

One of the most widely used databases for functional annotations is the Gene Ontology (GO) [2, 12]. This database is highly regarded both for its comprehensiveness and its unified approach for annotating genes in different species to the same basic set of underlying functions [2]. In order to evaluate the level of connection between a gene signature predicted by an experimental system and the set of genes that are annotated to a given

biological function, most functional enrichment analysis tools rely on set-overlap statistics [48]. Because these approaches are subject to an increase in type I errors associated with multiple hypothesis testing, corrections such as the Benjamini, Bonferroni, and FDR are often also applied [30] (discussed in more detail in Section 1.2.2).

Young et al. pointed out that these standard statistical approaches are sometimes inadequate, specifically when evaluating the functional properties of gene-sets derived from RNA-seq data since these experiments are prone to selection-bias due to variability in gene length [64]. Despite the wide use of functional analysis tools, however, little attention has been paid to whether or not the underlying properties of the functional databases themselves may contribute to spurious statistical results. For example, it is known that the number of annotations to functional terms in these databases has a heavy-tailed distribution [25]. To address this issue we investigate whether this heavy-tailed distribution leads to a bias in traditional functional analysis methods. We find that the significance level of the association between random gene sets and functions in the Gene Ontology appears to be positively correlated with the number of annotations made to the genes (degree of the genes) in a given gene set and the number of genes annotated to a particular GO category (degree of the GO term).

We investigate the properties of experimentally-

*contact: kglass@jimmy.harvard.edu

derived gene signatures, as reported in the Gene Signatures Database [15]. We focus on the annotation properties of the members of each gene signature, where these properties are defined based on the Gene Ontology. We find that most signatures include a disproportionate number of highly annotated genes. This is likely in part due to a non-independence between identified signatures and functional annotations, since genes that are involved in phenomena such as cancer are highly studied and thus more likely to be frequently annotated in databases such as GO. Given that these oncogenes are also likely to be identified in the experimental systems employed by researchers seeking to better improve our understanding of cancer, a method must be developed to account for the statistical bias surrounding gene sets with a larger than expected number of annotations.

Consequently, we propose a method, called Annotation Enrichment Analysis (AEA), that focuses on the overlap in *annotations* between a set of genes and the GO terms belonging to a particular branch of the GO hierarchy. By looking at annotation overlap instead of gene overlap our approach takes into account the annotation properties of the Gene Ontology. We then extend this concept and develop a structurally-preserving randomization scheme (SP-AEA) that evaluates the significance of annotation overlap by randomly sampling from the annotations contained in the Gene Ontology in order to even better capture the complex annotation relationships between genes and functional categories that may easily be missed by traditional set-overlap statistics. Implementations of both approaches are provided at <http://www.networks.umd.edu>.

1.2. Background

1.2.1. Properties of Functional Annotations

There are many functional annotation databases that have been developed in order to help classify genes according to their various roles in the cell [9, 29, 49, 50, 54]. These databases highlight different aspects of cellular function. Here, we focus our analysis on functional annotations made to the Gene Ontology because of its wide use by many functional enrichment tools (for example [1, 5, 28, 35, 53]). Since many of the annotation properties of the Gene Ontology are shared by other databases [25], we believe that the methods we develop here could be applied to functional enrichment analysis using other classification databases.

The Gene Ontology [2] takes the form of a directed acyclic graph (DAG) in which “child” functional categories (“terms”) can be subclassified under one or more other, more general categories, called “parent” terms, using “is a” and “part of” relationships. “Branches” in the Gene Ontology can therefore be defined as sets of terms that contain a parent term and all of its progeny. Note that these branches will contain overlapping sets of terms

since each term can be a descendant of multiple ancestors at each level of the DAG. Within this structure, genes are annotated to a set of functional categories related to that gene’s particular role in the cell. These annotations are transitive such that a parent term will take on all the genes annotations associated with any of its progeny [55]. Consequently, terms with many progeny often contain many gene annotations whereas terms with few progeny generally have fewer associated genes. “Biological Process,” “Molecular Function,” and “Cellular Component” are the three most general terms in GO, defining three independent branches such that every other term can only belong to one of these three categories. As a consequence all genes in GO are annotated to at least one, and often all three, of these categories. Note that since every parent term takes on the annotations of its progeny, the number of unique genes annotated to a parent term is the same as the number of unique genes annotated to the branch defined by the parent term and its progeny; however, the total number of annotations made to any parent term with children is less than the total number of annotations made to the corresponding branch (defined by the term and its progeny), since individual genes often have multiple annotations to terms within a branch.

Since we want to determine the influence of annotation properties on functional enrichment analysis, especially in the context of experimental gene signatures in commonly studied diseases such as cancer, we focus our study on annotations in the Gene Ontology associated with human genes. With this in mind we downloaded information regarding gene-term annotations for human genes from the Gene Ontology website (geneontology.org) and used this data to construct a gene-term bipartite graph, represented as an $n_G \times n_T$ adjacency matrix, where n_G is the total number of genes and n_T is the total number of terms listed in the annotation file. In this matrix a value of one indicates a known connection between the corresponding gene and term, and a value of zero indicates that the gene is not associated with that term. We will denote the $n_G \times n_T$ adjacency matrix of this bipartite graph by B . Thus

$$B_{ip} = \begin{cases} 1 & \text{if gene } i \text{ is annotated to term } p \\ 0 & \text{if gene } i \text{ is not annotated to term } p \end{cases} \quad (1)$$

Many terms are only associated with a small handful of genes, while some terms are associated with many genes.

A histogram of the “degree” of terms ($k_t^{(p)} = \sum_l^{n_G} B_{lp}$,

or the number of genes annotated to term p) reveals a heavy-tailed relationship (Figure 1(a)). In contrast, a

histogram of the “degree” of genes ($k_g^{(i)} = \sum_l^{n_T} B_{il}$, or the

number of terms to which gene i is annotated) shows that although some genes have many more annotations than others, a large portion of genes have approximately the same number of annotations (Figure 1(b)).

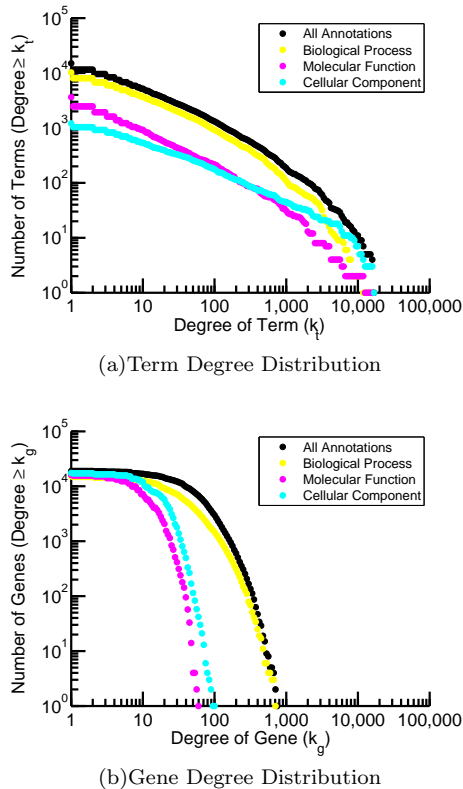


FIG. 1: The cumulative degree distributions of (a) genes and (b) terms in human GO annotations.

The “Biological Process” ontology contains a significant fraction of the total annotations. Although all three ontologies are used in functional enrichment analysis, it is common to focus on this ontology, both for its size and because its members describe dynamical processes performed by the cell. We will do the same in the following analysis. The total number of annotations made to the “Biological Process” ontology is 656783, originating from 18930 genes to 10192 terms. Consequently, the average number of annotations made by an individual gene is 43.2 and the average number of annotations made to an individual term is 64.4. These values will be useful to keep in mind, especially as we investigate the annotation properties of gene signatures and of the terms for which they are enriched.

1.2.2. Evaluating Functional Enrichment in Gene Sets using Set-Overlap Statistics

The most widely used statistics for evaluating which functional categories are enriched in a set of genes are based on gene counts and include Fisher’s Exact Test, the binomial test, and the chi-squared test [48]. Although these statistics vary in exact implementation, they all rely on the same basic underlying assumption that all genes have an equal probability of being selected under

the null hypothesis. Of these tests, Fisher’s Exact Test (FET) is the most common statistic and is used by many of the most popular functional enrichment tools (see Table 2 in [30]), and therefore we choose it to represent a “typical” evaluation of gene set functional enrichment. Although it is recognized that this statistic makes assumptions in its null hypothesis that fail to reflect the complex properties of the Gene Ontology, it is widely regarded as a good guide in determining what types of functions are represented in a given set of genes.

FET is related to the hypergeometric probability and can be used to calculate the significance, or p-value estimated using FET ($p_F(N_{gt})$) of an overlap between two independent sets. For example, given a gene set containing N_g genes, a GO term with k_t annotations, and N_{tot} total genes annotated in GO, the probability that N_{gt} or more genes belong both to this gene set and are annotated to the GO term can be calculated as:

$$p_F(N_{gt}) = P(N \geq N_{gt} | N_g, k_t, N_{tot}) = \sum_{i=N_{gt}}^{\min[N_g, k_t]} \frac{\binom{k_t}{i} \binom{N_{tot}-k_t}{N_g-i}}{\binom{N_{tot}}{N_g}}. \quad (2)$$

Since most functional enrichment analysis compares a gene set to all the terms in GO, multiple-hypothesis testing corrections are often applied to these p-values [30]. These corrections raise the value at which a comparison between a gene set and a GO term should be considered significant. Commonly used multiple-hypothesis corrections include the Bonferroni, Benjamini and the False Discovery Rate. Of these, the Bonferroni is the most conservative. It adjusts the value at which a test is considered “significant” by the number of tests made:

$$\beta = \alpha/n \quad (3)$$

where α is the value at which an individual test was previously considered significant, and β is the value at which an individual test should be considered significant, given n repetitions of that test. This is equivalent to multiplying the p-value obtained from a test such as FET by n and using the same critical value to determine whether the result is statistically significant.

The majority of multiple-hypothesis corrections will only change the critical value of individual tests, but will not affect the rank ordering of these tests. One popular exception to this is the False Discovery Rate (FDR). The FDR adjusts the value at which a test is considered “significant” based on the rank of the predicted level of significance:

$$\beta = \alpha n/r \quad (4)$$

where, as before α is the value at which an individual test was previously considered significant, β is the value at which an individual test should be considered significant, given n repetitions of that test, and r is the rank of the test. This will provide approximately the same

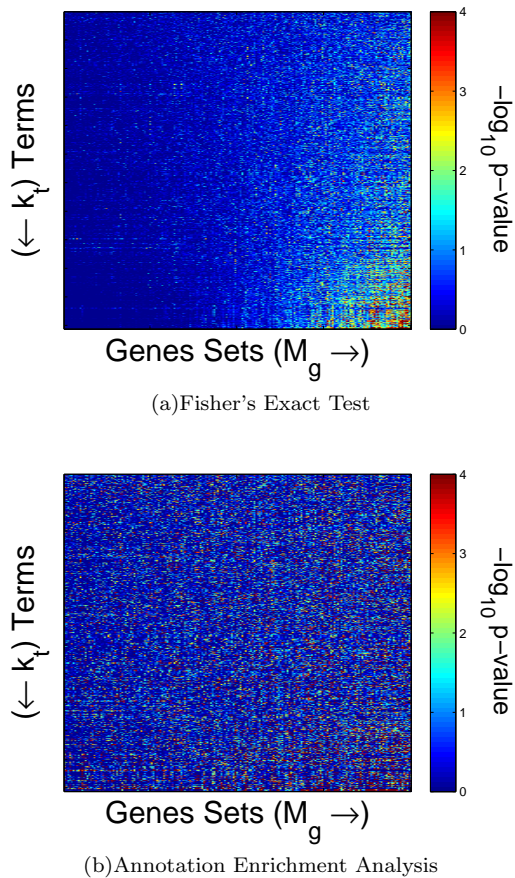


FIG. 2: The significance (measured by p-value) of GO terms in 200 randomly generated gene sets. The terms are ordered based on how many genes are annotated to the term (k_t) and the gene sets are ordered based on total the number of annotations (M_g) made by the 200 genes in that set. There is an obvious bias toward significant enrichment between high degree gene-set/term pairs in (a) Fisher’s Exact Test (FET), but this is correctly accounted for using (b) Annotation Enrichment Analysis (AEA).

correction as the Bonferroni for the most significantly-ranked p-values but will not adjust tests that are the least-significant by rank. As a consequence, the rank ordering of the significance can change slightly when using FDR.

2. METHODS

2.1. The Effect of Annotation Bias on Standard Functional Enrichment Analysis

To determine the effect of annotation properties on GO enrichment analysis we created random gene sets in which we controlled the total number of annotations made by the genes belonging to each gene set. Each set with a desired total number of annotations, M_g , was

created by first randomly selecting N_g genes. We then randomly selected one gene in this gene set (gene i) and one gene not in the gene set (gene j). If replacing gene i with gene j caused the total number of annotations made by genes in the gene set to approach the M_g , we replaced gene i with gene j with a high probability ($p = 0.95$), but if the replacement caused the average degree of the gene set to move farther away from M_g we replaced gene i with gene j with a low probability ($p = 0.05$). This swapping continued until the total number of annotations made by the gene set was within 0.1% of M_g .

In this way we created 200 gene sets with $N_g = 200$ genes each, but whose average degree ($k_{avg} = M_g/N_g$) varies from approximately 21 to 65, or from around half to 1.5 times the expected average degree of 43 (see Section 1.2.1). Using FET, we evaluated the enrichment of all GO terms in the “Biological Process” ontology in each of these random gene sets. Figure 2(a) shows a heat map of the significance of the overlap between each of these gene sets and GO terms that have 200 or more gene annotations, ordered based on their total number of gene annotations. The trend is striking. Gene sets with a higher number of annotations tend to have an abundance of enriched GO terms compared to gene sets with a lower number of annotations. Furthermore, GO terms with many gene annotations tend to be the most significantly enriched, especially in these “high-degree” gene sets. This indicates not only that false information is predicted for gene sets with many annotations, but that information may be lost for gene sets or GO categories with a relatively lower level of annotation.

We point out that although multiple-hypothesis corrections will sufficiently lower the value at which a p-value is considered significant such that either very few or no false positives will occur, the biases themselves cannot be overcome in this manner. Statistical corrections such as Bonferroni modify the threshold at which a result is considered significant (see Equation 3) by taking into account the number of different comparisons being made in a given statistical analysis, but do not consider the properties of the gene sets or terms themselves, thus using this correction the ordering of the p-values will not change and the bias will remain. Even an FDR correction (see Equation 4), which can alter the rank ordering of significance, is insufficient to overcome this strong signal (see Supplemental Figure 1).

2.2. Correcting for Annotation Bias using Annotation Enrichment Analysis

Clearly annotation properties of both genes and functional categories can influence the results of functional enrichment analysis. In order to mitigate this effect, we suggest that instead of considering the overlap between two gene sets, as is traditionally done in functional enrichment analysis, one instead considers the overlap between *annotations* made to a gene set and a branch of

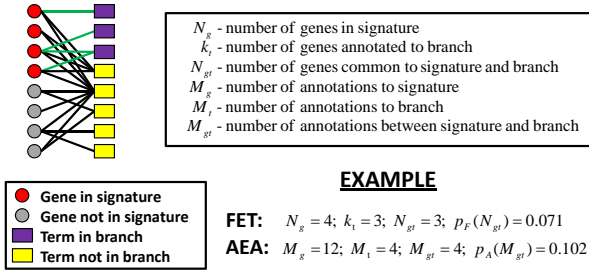


FIG. 3: An illustration of how AEA calculates the significance between a given gene signature and branch of the GO hierarchy (see Equation 5), and how that same information is calculated using FET (see Equation 2).

terms in the gene ontology. We call our method Annotation Enrichment Analysis (AEA).

Considering an entire branch of terms simultaneously is crucial to correctly estimating the annotation overlap between a gene set and a biological concept. As an illustration, consider the “Biological Process” category in GO, which is the ancestor of all other terms in the ontology and thus contains annotations from every gene. If we only considered the “Biological Process” term alone, then the number of annotations shared with this term and a gene set will be equal to the number of genes in the gene set, but it is more appropriate, given the nature of the term, for the number of common annotations to be equal to all the annotations made to the gene set. Using annotations to the full branch (a term and all its progeny) will correctly account for this.

Given that we want to estimate the significance of annotation overlap, one logical approach is to simply count the number of annotations made to a gene set, the number of annotations made to a branch in GO, and the number of annotations extending between that gene set and branch, and then once again use the hypergeometric probability to determine the significance of this overlap. This approach assumes that annotations are independent. Unfortunately, this is not the case for GO as a gene may only be annotated to a term a single time, but independent annotations would imply that the gene could be annotated to a term multiple times.

Acknowledging that we are making some false assumptions regarding the structure of gene-term annotations, we suggest that one can still estimate an enrichment for annotations between a gene set and a GO branch in the following manner. Given M_g annotations to a gene set, M_t annotations to terms belonging to a GO branch, and M_{tot} annotations made in the GO ontology, the probability of finding M_{gt} or more annotations in common

between these two sets can be written as:

$$p_A(M_{gt}) = P(M \geq M_{gt} | M_g, M_t, M_{tot}) = \sum_{i=M_{gt}}^{\min[M_g, M_t]} \frac{\binom{M_t}{i} \binom{M_{tot}-M_t}{M_g-i}}{\binom{M_{tot}}{M_g}}. \quad (5)$$

This equation can be used to calculate the significance (or p-value using AEA, $p_A(M_{gt})$) of enrichment between genes in a signature and genes annotated to a branch in the GO hierarchy, taking into account annotation information. A visual representation of how the significance of overlap would be calculated using FET compared to AEA is shown in Figure 3.

We determined the significance of all GO branches in our randomly generated gene sets with AEA, and created a heat map of these values as we had done for the significance values produced using standard set-overlap statistics (Figure 2(b)). The results of AEA are uniform across both varying gene set and term degree, demonstrating that AEA works well at eliminating annotation bias, although the predicted p-values are sometimes misleadingly low due to the independence assumption (for further discussion see Section 3.4).

3. RESULTS

3.1. Annotation Properties of Experimental Gene Signatures

One of the most common applications of GO enrichment analysis is to ascertain the functional properties of an experimentally determined set of genes. Although we have demonstrated that AEA corrects for annotation bias with randomly generated gene sets, we also want to know how well this analysis can recapitulate biologically-relevant results. With this in mind we downloaded signatures as recorded in Gene Signatures Database (GeneSigDB) [15]. This database is a manual curation of previously published gene expression signatures, focusing primarily on cancer and stem cell signatures [14]. In the following analysis we will use all 309 human signatures from this database that contain at least 100 and less than 1000 genes that also are annotated to a term in the “Biological Process” ontology.

First, to assess whether annotation bias might play a role in evaluating the functional properties of these gene signatures, we determined the average number of annotations made to the genes occurring in each signature. Figure 4(a) shows the number of genes in a signature plotted against the average level of annotation for each signature. The expectation for a random selection of genes (the average number of annotations made to all genes in GO – see Section 1.2.1) is shown as a red line. The plot suggests a strong preference for these signature genes to contain many more annotations than expected by chance. Almost a third (99) of the signatures have an average level

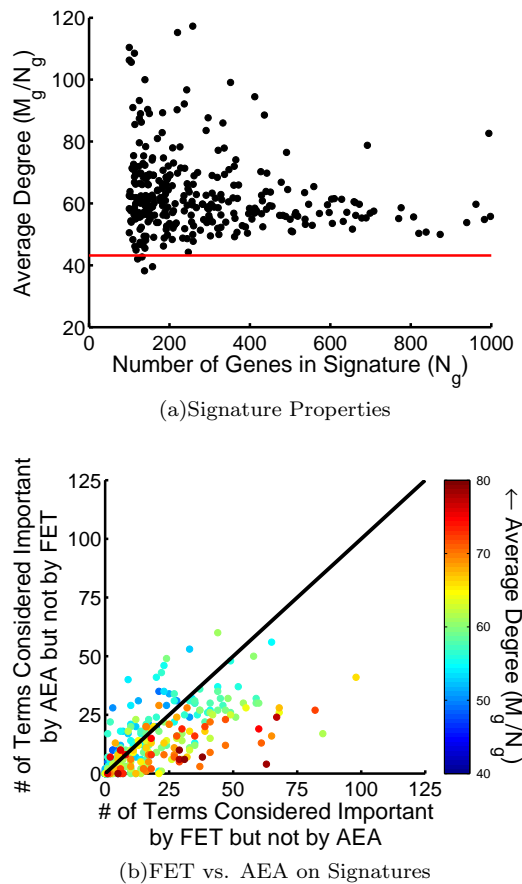


FIG. 4: Annotation properties of experimental gene signatures. (a) The number of genes versus the average number of annotations made to the genes in each signature. Genes from signatures generally contain many more GO annotations than one would expect if selecting genes randomly (shown in red). (b) The number of terms that are considered important (top 5% by rank) by one of the measures (either AEA or FET), but not important (bottom 85% by rank) by the other, plotted for each gene signature. The signatures are colored according to the average level of annotation ($k_{avg} = M_g/N_g$).

of annotation that is greater than any of our randomly generated gene sets (see Section 2.1) and all but four signatures have an average level of annotation greater than expected by chance. Since we have shown that random gene signatures with these annotation levels encounter a bias in traditional functional enrichment analysis, we believe these experimental signatures are an appropriate biological set with which to evaluate how AEA compares to FET when investigating and discovering the functions of genes contained in experimental biological data.

3.2. AEA vs FET on Expression Signatures

We predicted the enrichment of all “Biological Process” GO terms in these signatures both by traditional

set-overlap statistics (FET) as well as with AEA. We firstly investigated how the level of annotation made to a signature influences the results of the two metrics compared to one another. Although the exact functional categories considered most enriched according to the two measures may be different, we wanted to test if the two measures gave the same general results. In other words, are the categories ranked highly by FET also ranked highly by AEA and are the categories ranked poorly by FET also ranked poorly by AEA. To this end we selected the top 5% of terms (510 terms) based on their enrichment score in FET and AEA to designate as “important” according each to these measures. We compared this list of terms to the list of terms that are “not important” (in the bottom 85% of terms by rank) according to each measure. The number of terms considered important in AEA but not by FET versus the number of terms considered important by FET but not AEA for each signature is plotted in Figure 4(b). Signatures are colored based on the average level of annotation to their member genes.

In signatures containing the highest level of annotation, the terms deemed most “important” by FET are more likely to be considered “unimportant” according to AEA (note that “perfect” correlation between FET and AEA would result in a single point at (0,0)). Although the bias is not quite as strong, for signatures with lower levels of annotation, the terms deemed more “important” by AEA are sometimes considered “unimportant” by FET. These results are consistent with the previous analysis in random gene sets that showed a bias by FET to place more significance between gene sets and terms with a higher number of annotations (see Figure 2).

3.3. Specific Predictions made only by FET and not by AEA

To get a sense of whether the terms that are strongly enriched in signatures according to FET but not AEA constitute biologically meaningful information missed by AEA or uninformative predictions by FET we selected the ten term-signature pairs most significantly associated by Fisher’s exact test that are shown to be equivalent to chance ($p > 0.5$) by AEA. We also investigated the ten term-signature pairs most significantly associated by AEA that are given a p-value equivalent to chance ($p > 0.5$) by FET. These pairs are shown in Table 1 along with some of the annotation properties associated with those gene signatures and GO terms.

The selected pairs immediately suggest an FET bias for enrichment of signatures that include genes with a high number of annotations (as indicated by the large values in the k_{avg} column, expected value is approximately 60) as well as highly annotated GO terms (as indicated by the k_t column, expected value is 64.4). Furthermore, the terms that appear in the top ten most enriched pairs by FET and not AEA (e.g. “cellular process” and “metabolic process”) are quite general and reveal little information

Signature	GO category	k_{avg}	k_t	FET (FDR)	AEA (SP-AEA)
Breast (Supp. Table 1, [52])	cellular metabolic process	78.76	6259	2.24e-85 (1.16e-80)	0.99 (0.68)
Breast (Supp. Table 1, [52])	cellular macromolecule metabolic process	78.76	4382	1.84e-71 (5.09e-67)	1.00 (0.86)
Breast (Supp. Table 1, [52])	primary metabolic process	78.76	6541	3.66e-69 (9.00e-65)	1.00 (0.88)
Breast (Supp. Table 1, [52])	metabolic process	78.76	7234	8.47e-63 (1.72e-58)	1.00 (0.95)
ProteinKinases (TableS2, [34])	cellular metabolic process	86.00	6259	9.58e-55 (1.42e-50)	1.00 (1.00)
Breast (Supp. Table 1, [52])	macromolecule metabolic process	78.76	5088	5.46e-52 (6.99e-48)	1.00 (0.84)
ProteinKinases (TableS2, [34])	primary metabolic process	86.00	6541	7.86e-48 (8.14e-44)	1.00 (1.00)
StemCell (TableS3, [26])	cellular process	82.65	11520	8.41e-46 (7.98e-42)	1.00 (0.95)
StemCell (TableS3, [26])	nucleo-base, -side, -tide and nucleic acid metabolic process	82.65	2884	2.56e-43 (2.07e-39)	0.74 (0.60)
StemCell (TableS3, [26])	cellular nitrogen compound metabolic process	82.65	3301	2.42e-42 (1.82e-38)	1.00 (0.75)
Ovarian (Supp. Table 1, [3])	kidney development	55.00	128	0.63 (1.00)	8.11e-30 (0.014)
Breast (Supp. Table 1, [42])	kidney development	50.72	128	0.56 (1.00)	1.01e-14 (0.101)
Lymphoma (TableS4, [17])	kidney development	51.64	128	0.66 (1.00)	1.42e-13 (0.0505)
Breast (Supp. Table 1, [42])	renal system development	50.72	134	0.61 (1.00)	2.54e-13 (0.108)
Lymphoma (TableS4, [17])	renal system development	51.64	134	0.68 (1.00)	1.01e-12 (0.0554)
Stomach (TableS2b, [66])	cellular process	38.20	11520	0.66 (1.00)	1.45e-10 (0.0161)
Lymphoma (TableS4, [17])	urogenital system development	51.64	169	0.76 (1.00)	9.15e-10 (0.0664)
Lung (TableS2, [11])	regulation of cellular amine metabolic process	39.53	68	0.51 (1.00)	2.53e-08 (0.0414)
Breast (Supp. Table 6, [45])	regulation of T cell activation	49.39	211	0.57 (1.00)	3.15e-08 (0.0606)
Breast (Supp. Table 1, [19])	heart development	68.48	297	0.52 (1.00)	7.12e-08 (0.0845)

TABLE I: Term-signature pairs considered most significant by FET that are given a p-value equivalent to chance ($p > 0.5$) by AEA as well as pairs considered most significant by AEA that are given a p-value equivalent to chance by FET. Signatures are identified based on their cell-type, publication, and table reference, as reported in GeneSigDB. For the pairs enriched in FET and not AEA, the average degree of the genes in these signatures (k_{avg}) and the degree of the terms they are enriched for (k_t) both tend to be of higher degree (even compared to the already high degree of signatures considered in this analysis) and it is unclear how the functions uncovered by FET and not AEA are important for the specific biological systems represented by these signatures. In contrast, in the pairs selected as enriched by AEA and not FET, the average degree of the genes in the signatures (k_{avg}) and the degree of the terms they are enriched for (k_t) are much closer to expectation than those that are significant by FET, however, the assuming independence between annotations led to several mis-leadingly low p-values where a single highly annotated gene is able to drastically effect the estimated significance.

about the *specific* biology performed by the genes in the signature. Based on this list it appears that the results from FET that are discounted by AEA are generally uninformative. Although many of the genes in these signatures indeed perform the listed biological functions, it is likely that the enrichment reported by FET is a consequence of the fact that the genes in the signature perform many functions. It is useful to point out that these “spurious results” have, by no means, questionable p-values. Even with multiple-hypothesis corrections (FDR noted in parentheses) these values remain small enough to conclude that these gene signatures are involved in these particular biological functions, when in reality, this “significant” level of association may merely be due to the annotation properties of the signatures and terms themselves.

In contrast, the term-signature pairs enriched in AEA and not FET are no longer biased toward high-degree signatures and are not as strongly biased toward high-degree terms. The enriched terms selected by AEA and not FET are also much more “specific” compared to the ones selected by FET and not AEA and may at first appear intriguing given cancer’s identified connection with stem cells (see for example, [43]), and hence developmental processes. Further investigation, however, shows that the strong enrichment values are largely a consequence

of the invalid assumption made in Equation 5 that annotations are independent of genes. This is manifested in the fact that many of these term-signature pairs share many annotations in common, but only one or two genes. For example, the genes contained in the top signature enriched in AEA and not FET (Ovarian (Supp. Table 1, [3])) collectively include a total of 6380 annotations, of which 61 extend to either “kidney development” or one of its progeny, a branch of GO containing a total of 966 annotations. This is a very significant overlap ($8.11e - 30$), given the 656783 total annotations made to terms in the “Biological Process” ontology; however, all 61 annotations originate from the same highly-annotated gene, OSR1 (263 total annotations in the “Biological Process” ontology). Because we had assumed that annotations are independent of one another, the information that an annotation to OSR1 normally is connected to many other annotations, was lost when calculating the p-value. In other words, Equation 5 effectively ignores information regarding the degree of individual genes and terms, thus allowing a single high-degree gene in a signature to incorrectly influence the resulting p-values.

3.4. A Structure Preserving Form of Annotation Enrichment Analysis

Although AEA both successfully corrects for annotation bias and is conceptually appealing, it assumes that annotations made to genes and terms are completely independent of one another, ignoring much of the structure contained in the Gene Ontology. In this section we develop a randomization scheme that preserves the structure of GO annotations while calculating the significance the number of co-annotations between a gene set and a GO branch. This scheme preserves the annotation properties of individual genes and terms, thus incorporating that structure into the background distribution and allowing us to even more accurately assess the significance of overlap between a given gene signature and GO branch. We call this approach structure-preserving AEA (SP-AEA) and illustrate it in Figure 5.

In this randomization is it again useful to think of the Gene Ontology as a bipartite graph (see Section 1.2.1). As with AEA, for SP-AEA we begin by determining M_g , the number of annotations to a gene set, M_t , the number of annotations to the terms in a GO branch, and M_{gt} , the number of annotations stretching between this gene set and branch. We then determine the expected number of co-annotations between this branch and a random sets of genes or between this set of genes and a random selection of terms. In order to do this, we perform two randomizations, one of the order of genes and the other of the order of terms, while still preserving the original connections from the GO bipartite graph. We then take annotations connected to the top random genes until we've selected M_g annotations, and determine \tilde{M}_g , the number of edges in the bipartite graph that extend between the top random genes and the branch of the GO hierarchy. Similarly, we take annotations connected to the top random terms until we've selected M_t annotations, and determine \tilde{M}_t , the number of edges in the bipartite graph that extend between the top random terms and the original gene set. In the (fairly common) case where selecting the top M_g/M_t annotations does not correspond to selecting a whole number of genes/terms, we take a random selection of edges from the final gene/term selected in order to get exactly M_g/M_t annotations. We repeat the randomization process many times in order to determine a distribution of values for \tilde{M}_g given the gene annotations, and a distribution of values for \tilde{M}_t given term annotations. In order to summarize the result, we define a new p-value, $p_S(M_{gt})$ which reflects the significance of the M_{gt} annotations by averaging the probability that $\tilde{M}_g \geq M_{gt}$ and the probability that $\tilde{M}_t \geq M_{gt}$:

$$p_S(M_{gt}) \equiv \frac{1}{2}P(\tilde{M}_g \geq M_{gt}) + \frac{1}{2}P(\tilde{M}_t \geq M_{gt}). \quad (6)$$

We tested the performance of this approximation (using 10,000 randomizations) by determining the functional enrichment of GO terms in our randomly gener-

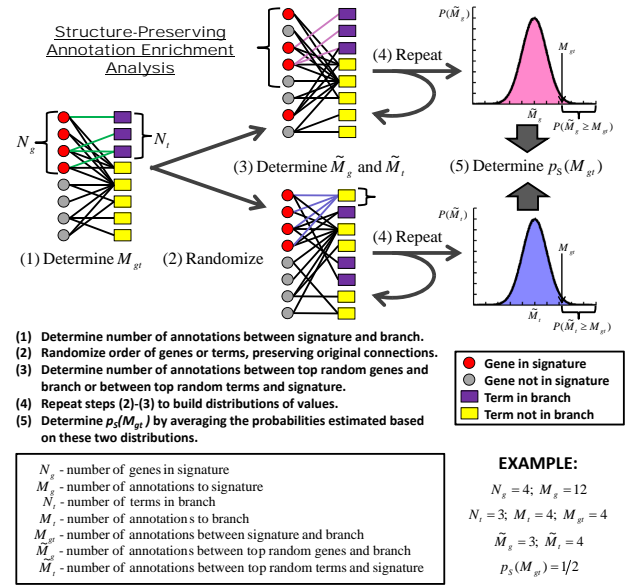


FIG. 5: An outline of how Structure-Preserving Annotation Enrichment Analysis (SP-AEA) calculates the significance of association between a given gene signature and the collection of terms that belong to a branch in the GO hierarchy.

ated gene sets (see Section 2.1). As with AEA, SP-AEA effectively eliminates annotation bias (see Supplemental Figure 2). Further, the predicted p-values are much more reasonable compared to those made by either FET or AEA as only 9 gene-set/term pairs have a p-value less than 10^{-4} according to SP-AEA compared to 433 in FET and over two million in AEA (compare Supplemental Figure 2 with Figure 2). This indicates that SP-AEA both correctly accounts for annotation bias between various gene sets and terms, *and* also recognizes that these are indeed random selections of genes that should have no coordinated functional properties.

Finally, we ran SP-AEA on our experimental gene signatures (one million randomizations). In order to verify that preserving the structure of GO annotations in SP-AEA adequately negates the influence of single genes on p-values, we determined the top ten term-signature pairs enriched by FET and not by SP-AEA and vice versa. The pairs enriched in FET and not SP-AEA are the exact same pairs that were enriched in FET and not AEA (see Table 1), however, no signature-term pairs are enriched in SP-AEA at a significance less than 0.01 (and only three with a significance less than 0.05) that have a p-value greater than 0.5 by FET. Similarly, no pairs are predicted as significant by SP-AEA that are not also significant by AEA. P-values as predicted by SP-AEA are shown for the term-signature pairs in Table 1.

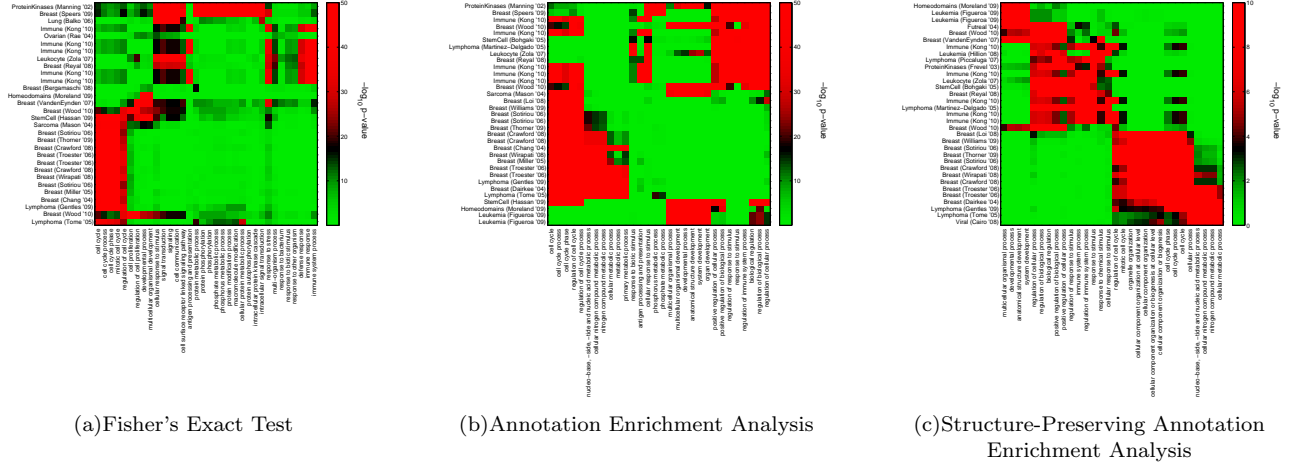


FIG. 6: Clustergrams representing enriched term-signature pairs. (a) A clustering of signatures and terms selected based on their enrichment-score according to FET. These signatures (from [4, 6, 10, 13, 24, 26, 34, 37–40, 46, 47, 51, 52, 56, 57, 59, 62, 63, 67]) did not cluster as well into distinct biologically units. (b) A clustering of signatures and terms selected based on their enrichment-score according to SP-AEA. These signatures look like they may fall into three clusters generally associated with regulation and stimulus response (from [7, 34, 36, 38, 47, 52, 63, 67]), cycle cycle and metabolic processes (from [10, 13, 16, 24, 33, 37, 39, 51, 56–58, 61, 62]) and developmental processes (from [21, 26, 40]). This clustering is something not apparent using the FET measure but even more apparent using SP-AEA. (c) A clustering of signatures and terms selected based on their enrichment-score according to SP-AEA. The signatures and terms break into several, biologically distinct units. One includes cellular-differentiation signatures published in [21, 23, 40, 59]. The second is associated with immune-response, and includes signatures published in [7, 22, 27, 36, 38, 44, 47, 63, 67]. Another cluster includes breast cancer signatures published in [13, 16, 33, 51, 56, 58, 61, 62]. Finally two lymphoma [24, 57] and a viral signature [8] associated with proliferation are also included.

3.5. AEA Reveals Biologically-Relevant Functions

Finally, we investigated the kind of biology that is highlighted using AEA and SP-AEA compared to FET. In order to focus on the most relevant biology we selected 30 representative terms/signatures for each measure. We selected these terms/signatures by ranking according to the minimum enrichment each has across all signatures/terms and selecting top 30 by this rank. For SP-AEA, since many (586) term-signature pairs have an estimated p-value of zero even after one million randomizations, we broke ties by the number of signatures/terms enriched in the terms/signatures at this level. We then performed hierarchical clustering (using the “clustergram” function in Matlab using default settings) on the terms and signatures selected for each of the three measures. The results are shown in Figure 6.

Clustering the FET results does not give rise to any apparent term-signature structure, however, several of the individual pairs highlight important biological processes. For example, the FET clustering shows an enrichment of cell-cycle related processes in breast cancer signatures [32] and includes immune-related terms enriched in immune gene signatures. This, however, accounts for only a handful of the selected terms; the clustergram also includes a number of functional categories related to “proteins” and “phosphorylation” that are only enriched in

a small number of signatures. Although this analysis should not be considered evidence that FET produces incorrect or unmeaningful results, we do suggest that the results presented in the clustergram and as well as in the preceding sections indicate that the results of FET can be, and often are, muddled by a signal driven by annotation bias, highlighting either highly-annotated signatures or more general biological processes that may not be specific to the systems in question.

This becomes evident in the clustergrams of the terms and signatures selected using either AEA or SP-AEA where much of the biology, although often present in FET, is more obvious because of the removal of the annotation bias. The clearest results are obtained using the structure-preserving SP-AEA where several distinct clusters of signatures and terms emerge that are correlated with categories such as cellular-differentiation markers, immune-response, and breast cancer.

The terms associated with each cluster represent biological processes that are known to be involved in these types of signatures. For example the cluster enriched for terms such as “system development” and “developmental process” includes signatures identified based on their role in cellular differentiation. A signature associated with homeodomains is included in this cluster. This is consistent with the biological role of homeodomains, since many are known to initiate cascades of genes, which

in turn induces cellular differentiation into tissues and organs. The Leukemia signatures [21] were both obtained by looking at the expression levels from cells where CEBP α , a protein known to induced differentiation in Leukemia cells [65], is silenced. The cluster also includes a set of oncogenes [23], as well as two breast cancer signatures, one of which includes genes involved in angiogenesis [59].

In contrast, the cluster that primarily includes signatures from immune-systems, lymphoma and leucocytes, is logically also enriched in terms such as “immune system” and “response to stimulus” as well as terms related to “biological regulation”. Interestingly, the breast signature strongly associated with this cluster [47] represents a list of genes defined based on immune response in breast cancer and the stem cell signature [7] is from a study on patients with systemic sclerosis, a type of autoimmune disorder. In addition, the protein-kinase signature [22] included in the cluster is from a study of p58 Mitogen-activated protein (MAP) kinase signaling of mRNA stability. MAP kinases have been shown to play an important role in immune response [20].

Another cluster associated only with breast cancer signatures shows a strong enrichment for terms related to the cell cycle and cellular component organization, processes known to be differentially regulated in breast cancer [32]. Finally, two lymphoma and one viral signature that were identified based on cell proliferation (for example, by association with Myc targeting [8, 24]) are enriched for terms such as “cellular metabolic process.” This is consistent with expectation since there is evidence that a connection exists between proliferation and metabolic pathways in cancer cells [18, 60].

4. CONCLUSION

By considering annotation properties, AEA is able to highlight the roles of genes in previously published gene

signatures that, although found by FET, were obscured. The limitations of FET in this context is largely attributable to the fact that many of these published gene-signatures include a large number of highly-annotated genes that, as a consequence, influences the results of functional enrichment analysis. Of course, the association between highly annotated genes and their more abundant occurrence in published gene sets may be at least partially attributable to the fact that some of these same publications may have been used in assigning a subset of the annotations in GO. We point out that although it is possible that newly-derived gene signatures may not exhibit the same level of annotation-bias as previously-published signatures, it is also very probable that highly annotated genes are important in a wide variety of well-studied systems and will continue to show up and influence the results of functional enrichment analysis on newly generated gene sets. In light of this we suggest using our approach alongside or in place of other traditional measures, especially for gene signatures that are known to contain significantly more or less annotations than one would expect by chance. We believe that AEA will allow biologists to better interpret the functional roles of genes identified as important in their experimental system.

Acknowledgement

We would like to thank Emanuele Mazzola for helpful discussions regarding this work.

-
- [1] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, July 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl140. URL <http://dx.doi.org/10.1093/bioinformatics/btl140>.
 - [2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics*, 25(1): 25–29, May 2000. ISSN 1061-4036. doi: 10.1038/75556. URL <http://dx.doi.org/10.1038/75556>.
 - [3] Dimcho Bachvarov, Sylvain L’esperance, Ion Popa, Magdalena Bachvarova, Marie Plante, and Bernard Têtu. Gene expression patterns of chemoresistant and chemosensitive serous epithelial ovarian tumors with possible predictive value in response to initial chemotherapy. *International journal of oncology*, 29(4):919–933, October 2006. ISSN 1019-6439. URL <http://view.ncbi.nlm.nih.gov/pubmed/16964388>.
 - [4] Justin M. Balko, Anil Potti, Christopher Saunders, Arnold Stromberg, Eric B. Haura, and Esther P. Black. Gene expression patterns that predict sensitivity to epidermal growth factor receptor tyrosine kinase inhibitors in lung cancer cell lines and human lung tumors. *BMC Genomics*, 7:289+, November 2006. ISSN 1471-2164. doi: 10.1186/1471-2164-7-289. URL <http://dx.doi.org/10.1186/1471-2164-7-289>.
 - [5] Tim Beissbarth and Terence P. Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics (Oxford, England)*, 20

- (9):1464–1465, June 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth088. URL <http://dx.doi.org/10.1093/bioinformatics/bth088>.
- [6] A. Bergamaschi, E. Tagliabue, T. Sørli, B. Naume, T. Triulzi, R. Orlandi, H. G. Russnes, J. M. Nesland, R. Tammi, P. Auvinen, V-M M. Kosma, S. Ménard, and A-L L. Børresen-Dale. Extracellular matrix signature identifies breast cancer subgroups with different clinical outcome. *The Journal of pathology*, 214(3):357–367, February 2008. ISSN 0022-3417. doi: 10.1002/path.2278. URL <http://dx.doi.org/10.1002/path.2278>.
- [7] T. Bohgaki, Y. Amasaki, N. Nishimura, M. Bohgaki, Y. Yamashita, M. Nishio, K-I I. Sawada, S. Jodo, T. Atsumi, and T. Koike. Up regulated expression of tumour necrosis factor alpha converting enzyme in peripheral monocytes of patients with early systemic sclerosis. *Annals of the rheumatic diseases*, 64(8):1165–1173, August 2005. ISSN 0003-4967. doi: 10.1136/ard.2004.030338. URL <http://dx.doi.org/10.1136/ard.2004.030338>.
- [8] Stefano Cairo, Carolina Armengol, Aurélien De Reynies, Yu Wei, Emilie Thomas, Claire-Angélique A. Renard, Andrei Goga, Asha Balakrishnan, Michaela Semeraro, Lionel Gresh, Marco Pontoglio, Hélène Strick-Marchand, Florence Levillayer, Yann Nouet, David Rickman, Frédéric Gauthier, Sophie Branchereau, Laurence Brugières, Véronique Laithier, Raymonde Bouvier, Françoise Boman, Giuseppe Basso, Jean-François F. Michiels, Paul Hofman, Francine Arbez-Gindre, Hélène Jouan, Marie-Christine C. Rousselet-Chapeau, Dominique Berrebi, Luc Marcellin, François Plenat, Dominique Zachar, Madeleine Joubert, Janick Selves, Dominique Pasquier, Paulette Bioulac-Sage, Michael Grotzer, Margaret Childs, Monique Fabre, and Marie-Annick A. Buendia. Hepatic stem-like phenotype and interplay of wnt/beta-catenin and myc signaling in aggressive childhood liver cancer. *Cancer cell*, 14(6):471–484, December 2008. ISSN 1878-3686. doi: 10.1016/j.ccr.2008.11.002. URL <http://dx.doi.org/10.1016/j.ccr.2008.11.002>.
- [9] Ron Caspi, Tomer Altman, Joseph M. Dale, Kate Dreher, Carol A. Fulcher, Fred Gilham, Pallavi Kaipa, Athikattuvalasu S. Karthikeyan, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Suzanne Paley, Liviu Popescu, Anuradha Pujar, Alexander G. Shearer, Peifen Zhang, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, 38(Database issue): D473–D479, January 2010. ISSN 1362-4962. doi: 10.1093/nar/gkp875. URL <http://dx.doi.org/10.1093/nar/gkp875>.
- [10] Howard Y. Chang, Julie B. Sneddon, Ash A. Alizadeh, Ruchira Sood, Rob B. West, Kelli Montgomery, Jen-Tsan T. Chi, Matt van de Rijn, David Botstein, and Patrick O. Brown. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS biology*, 2(2):e7+, February 2004. ISSN 1545-7885. doi: 10.1371/journal.pbio.0020007. URL <http://dx.doi.org/10.1371/journal.pbio.0020007>.
- [11] Christopher D. Coldren, Barbara A. Helfrich, Samir E. Witta, Michio Sugita, Razvan Lapadat, Chan Zeng, Anna Barón, Wilbur A. Franklin, Fred R. Hirsch, Mark W. Geraci, and Paul A. Bunn. Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines. *Molecular cancer research : MCR*, 4(8):521–528, August 2006. ISSN 1541-7786. doi: 10.1158/1541-7786.MCR-06-0095. URL <http://dx.doi.org/10.1158/1541-7786.MCR-06-0095>.
- [12] The Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Research*, 38(suppl 1):D331–D335, January 2010. ISSN 1362-4962. doi: 10.1093/nar/gkp1018. URL <http://dx.doi.org/10.1093/nar/gkp1018>.
- [13] Nigel P. Crawford, Jude Alsarraj, Luanne Lukes, Renard C. Walker, Jennifer S. Officewala, Howard H. Yang, Maxwell P. Lee, Keiko Ozato, and Kent W. Hunter. Bromodomain 4 activation predicts breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America*, 105(17):6380–6385, April 2008. ISSN 1091-6490. doi: 10.1073/pnas.0710331105. URL <http://dx.doi.org/10.1073/pnas.0710331105>.
- [14] Aedín C. Culhane, Thomas Schwarzl, Razvan Sultana, Kermshlise C. Picard, Shaita C. Picard, Tim H. Lu, Katherine R. Franklin, Simon J. French, Gerald Papenhausen, Mick Correll, and John Quackenbush. GeneSigDB—a curated database of gene expression signatures. *Nucleic acids research*, 38(Database issue):D716–D725, January 2010. ISSN 1362-4962. doi: 10.1093/nar/gkp1015. URL <http://dx.doi.org/10.1093/nar/gkp1015>.
- [15] Aedín C. Culhane, Markus S. Schröder, Razvan Sultana, Shaita C. Picard, Enzo N. Martinelli, Caroline Kelly, Benjamin Haibe-Kains, Misha Kapushesky, Anne-Alyssa St Pierre, William Flahive, Kermshlise C. Picard, Daniel Gusenleitner, Gerald Papenhausen, Niall O’Connor, Mick Correll, and John Quackenbush. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Research*, 40(D1):D1060–D1066, January 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr901. URL <http://dx.doi.org/10.1093/nar/gkr901>.
- [16] Shanaz H. Dairkee, Youngran Ji, Yong Ben, Dan H. Moore, Zhenhang Meng, and Stefanie S. Jeffrey. A molecular ‘signature’ of primary breast cancer cultures; patterns resembling tumor tissue. *BMC genomics*, 5(1), July 2004. ISSN 1471-2164. doi: 10.1186/1471-2164-5-47. URL <http://dx.doi.org/10.1186/1471-2164-5-47>.
- [17] Laurence de Leval, David S. Rickman, Caroline Thielen, Aurélien de Reynies, Yen-Lin L. Huang, Georges Delsol, Laurence Lamant, Karen Leroy, Josette Brière, Thierry Molina, Françoise Berger, Christian Gisselbrecht, Luc Xerri, and Philippe Gaulard. The gene expression profile of nodal peripheral t-cell lymphoma demonstrates a molecular link between angioimmunoblastic t-cell lymphoma (AITL) and follicular helper t (TFH) cells. *Blood*, 109(11):4952–4963, June 2007. ISSN 0006-4971. doi: 10.1182/blood-2006-10-055145. URL <http://dx.doi.org/10.1182/blood-2006-10-055145>.
- [18] Ralph J. DeBerardinis, Julian J. Lum, Georgia Hatzivassiliou, and Craig B. Thompson. The biology of cancer: Metabolic reprogramming fuels cell growth and proliferation. *Cell Metabolism*, 7(1):11–20, January 2008. ISSN 15504131. doi: 10.1016/j.cmet.2007.10.002. URL <http://dx.doi.org/10.1016/j.cmet.2007.10.002>.
- [19] Christine Desmedt, Benjamin Haibe-Kains, Pratyaksha Wirapati, Marc Buyse, Denis Larsimont, Gianluca Bontempì, Mauro Delorenzi, Martine Piccart, and Chris-

- tos Sotiriou. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14(16):5158–5165, August 2008. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-07-4756. URL <http://dx.doi.org/10.1158/1078-0432.CCR-07-4756>.
- [20] Chen Dong, Roger J. Davis, and Richard A. Flavell. MAP kinases in the immune response. *Annual review of immunology*, 20:55–72, 2002. ISSN 0732-0582. doi: 10.1146/annurev.immunol.20.091301.131133. URL <http://dx.doi.org/10.1146/annurev.immunol.20.091301.131133>.
- [21] Maria E. Figueroa, Bas J. Wouters, Lucy Skrabanek, Jacob Glass, Yushan Li, Claudia A. Erpelinck-Verschueren, Anton W. Langerak, Bob Löwenberg, Melissa Fazzari, John M. Greally, Peter J. Valk, Ari Melnick, and Ruud Delwel. Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. *Blood*, 113(12):2795–2804, March 2009. ISSN 1528-0020. doi: 10.1182/blood-2008-08-172387. URL <http://dx.doi.org/10.1182/blood-2008-08-172387>.
- [22] Mathias A. Frevel, Tala Bakheet, Aristobolo M. Silva, John G. Hissong, Khalid S. Khabar, and Bryan R. Williams. p38 mitogen-activated protein kinase-dependent and -independent signaling of mRNA stability of AU-rich element-containing transcripts. *Molecular and cellular biology*, 23(2):425–436, January 2003. ISSN 0270-7306. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC151534/>.
- [23] P. Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. A census of human cancer genes. *Nature reviews. Cancer*, 4(3):177–183, March 2004. ISSN 1474-175X. doi: 10.1038/nrc1299. URL <http://dx.doi.org/10.1038/nrc1299>.
- [24] Andrew J. Gentles, Ash A. Alizadeh, Su-In I. Lee, June H. Myklebust, Catherine M. Shachaf, Babak Shahbaba, Ronald Levy, Daphne Koller, and Sylvia K. Plevritis. A pluripotency signature predicts histologic transformation and influences survival in follicular lymphoma patients. *Blood*, 114(15):3158–3166, October 2009. ISSN 1528-0020. doi: 10.1182/blood-2009-02-202465. URL <http://dx.doi.org/10.1182/blood-2009-02-202465>.
- [25] Kimberly Glass, Edward Ott, Wolfgang Losert, and Michelle Girvan. Implications of functional similarity for gene regulatory interactions. *Journal of the Royal Society, Interface / the Royal Society*, February 2012. ISSN 1742-5662. doi: 10.1098/rsif.2011.0585. URL <http://dx.doi.org/10.1098/rsif.2011.0585>.
- [26] Khaled A. Hassan, Guoan Chen, Gregory P. Kalemkerian, Max S. Wicha, and David G. Beer. An embryonic stem cell-like signature identifies poorly differentiated lung adenocarcinoma but not squamous cell carcinoma. *Clinical Cancer Research*, 15(20):6386–6390, October 2009. doi: 10.1158/1078-0432.CCR-09-1105. URL <http://dx.doi.org/10.1158/1078-0432.CCR-09-1105>.
- [27] Joelle Hillion, Surajit Dhara, Takita Felder F. Sumter, Mita Mukherjee, Francescopaolo Di Cello, Amy Belton, James Turkson, Souyma Jaganathan, Linzhao Cheng, Zhaohui Ye, Richard Jove, Peter Aplan, Ying-Wei W. Lin, Kelsey Wertzler, Ray Reeves, Ossama Elbahlouh, Jeanne Kowalski, Raka Bhattacharya, and Linda M. Resar. The high-mobility group a1a/signal transducer and activator of transcription-3 axis: an achilles heel for hematopoietic malignancies? *Cancer research*, 68(24):10121–10127, December 2008. ISSN 1538-7445. doi: 10.1158/0008-5472.CAN-08-2121. URL <http://dx.doi.org/10.1158/0008-5472.CAN-08-2121>.
- [28] Da W. Huang, Brad T. Sherman, Qina Tan, Joseph Kir, David Liu, David Bryant, Yongjian Guo, Robert Stephens, Michael W. Baseler, H. Clifford Lane, and Richard A. Lempicki. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucl. Acids Res.*, 35(Web Server issue):gkm415+, June 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm415. URL <http://dx.doi.org/10.1093/nar/gkm415>.
- [29] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.27. URL <http://dx.doi.org/10.1093/nar/28.1.27>.
- [30] Purvesh Khatri and Sorin Drăghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, September 2005. ISSN 1460-2059. doi: 10.1093/bioinformatics/bti565. URL <http://dx.doi.org/10.1093/bioinformatics/bti565>.
- [31] O. D. King, R. E. Foulger, S. S. Dwight, J. V. White, and F. P. Roth. Predicting gene function from patterns of annotation. *Genome research*, 13(5):896–904, May 2003. ISSN 1088-9051. doi: 10.1101/gr.440803. URL <http://dx.doi.org/10.1101/gr.440803>.
- [32] M. Loddo, S. R. Kingsbury, M. Rashid, I. Proctor, C. Holt, J. Young, S. El-Sheikh, M. Falzon, K. L. Eward, T. Prevost, R. Sainsbury, K. Stoeber, and G. H. Williams. Cell-cycle-phase progression analysis identifies unique phenotypes of major prognostic and predictive significance in breast cancer. *British Journal of Cancer*, aop(current). ISSN 0007-0920. doi: 10.1038/sj.bjc.6604924. URL <http://dx.doi.org/10.1038/sj.bjc.6604924>.
- [33] Sherene Loi, Benjamin Haibe-Kains, Christine Desmedt, Pratyaksha Wirapati, Françoise Lallemand, Andrew M. Tutt, Cheryl Gillet, Paul Ellis, Kenneth Ryder, James F. Reid, Maria G. Daidone, Marco A. Pierotti, Els Mjj M. Berns, Maurice Phm P. Jansen, John A. Foekens, Mauro Delorenzi, Gianluca Bontempi, Martine J. Piccart, and Christos Sotiriou. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC genomics*, 9(1):239+, May 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-239. URL <http://dx.doi.org/10.1186/1471-2164-9-239>.
- [34] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, December 2002. ISSN 1095-9203. doi: 10.1126/science.1075762. URL <http://dx.doi.org/10.1126/science.1075762>.
- [35] David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. GOToolBox: functional analysis of gene datasets based on gene ontology. *Genome Biol*, 5(12), 2004. ISSN 1465-6914. doi: 10.1186/gb-2004-5-12-r101. URL <http://dx.doi.org/10.1186/gb-2004-5-12-r101>.
- [36] B. Martínez-Delgado, M. Cuadros, E. Honrado, A. Ruiz de la Parte, G. Roncador, J. Alves, J. M. Castrillo,

- C. Rivas, and J. Benítez. Differential expression of NF-kappaB pathway genes among peripheral t-cell lymphomas. *Leukemia*, 19(12):2254–2263, December 2005. ISSN 0887-6924. doi: 10.1038/sj.leu.2403960. URL <http://dx.doi.org/10.1038/sj.leu.2403960>.
- [37] Douglas X. Mason, Tonya J. Jackson, and Athena W. Lin. Molecular signature of oncogenic ras-induced senescence. *Oncogene*, 23(57):9238–9246, December 2004. ISSN 0950-9232. doi: 10.1038/sj.onc.1208172. URL <http://dx.doi.org/10.1038/sj.onc.1208172>.
- [38] Y. Megan Kong, Carl Dahlke, Qun Xiang, Yu Qian, David Karp, and Richard H. Scheuermann. Toward an ontology-based framework for clinical research databases. *Journal of biomedical informatics*, May 2010. ISSN 1532-0480. doi: 10.1016/j.jbi.2010.05.001. URL <http://dx.doi.org/10.1016/j.jbi.2010.05.001>.
- [39] Lance D. Miller, Johanna Smeds, Joshy George, Vinsensius B. Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan, Per Hall, Sigrid Klaar, Edison T. Liu, and Jonas Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13550–13555, September 2005. ISSN 1091-6490. doi: 10.1073/pnas.0506230102. URL <http://dx.doi.org/10.1073/pnas.0506230102>.
- [40] R. Travis Moreland, Joseph F. Ryan, Christopher Pan, and Andreas D. Baxevanis. The homeodomain resource: a comprehensive collection of sequence, structure, interaction, genomic and functional information on the homeodomain protein family. *Database : the journal of biological databases and curation*, 2009, 2009. ISSN 1758-0463. doi: 10.1093/database/bap004. URL <http://dx.doi.org/10.1093/database/bap004>.
- [41] Sara Mostafavi and Quaid Morris. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics (Oxford, England)*, 26(14):1759–1765, July 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq262. URL <http://dx.doi.org/10.1093/bioinformatics/btq262>.
- [42] Ankur K. Nagaraja, Chad J. Creighton, Zhifeng Yu, Huifeng Zhu, Preethi H. Gunaratne, Jeffrey G. Reid, Emuejevoke Olokpa, Hiroaki Itamochi, Naoto T. Ueno, Shannon M. Hawkins, Matthew L. Anderson, and Martin M. Matzuk. A link between mir-100 and FRAP1/mTOR in clear cell ovarian cancer. *Molecular endocrinology (Baltimore, Md.)*, 24(2):447–463, February 2010. ISSN 1944-9917. doi: 10.1210/me.2009-0295. URL <http://dx.doi.org/10.1210/me.2009-0295>.
- [43] Jae-Il I. Park, Andrew S. Venteicher, Ji Yeon Y. Hong, Jinkuk Choi, Sohee Jun, Marina Shkreli, Woody Chang, Zhaojing Meng, Peggie Cheung, Hong Ji, Margaret McLaughlin, Timothy D. Veenstra, Roel Nusse, Pierre D. McCrea, and Steven E. Artandi. Telomerase modulates wnt signalling by association with target gene chromatin. *Nature*, 460(7251):66–72, July 2009. ISSN 1476-4687. doi: 10.1038/nature08137. URL <http://dx.doi.org/10.1038/nature08137>.
- [44] Pier Paolo P. Piccaluga, Claudio Agostinelli, Andrea Califano, Maura Rossi, Katia Basso, Simonetta Zupo, Philip Went, Ulf Klein, Pier Luigi L. Zinzani, Michele Bacarani, Riccardo Dalla Favera, and Stefano A. Pileri. Gene expression analysis of peripheral t cell lymphoma, unspecified, reveals distinct profiles and new potential therapeutic targets. *The Journal of clinical investigation*, 117(3):823–834, March 2007. ISSN 0021-9738. doi: 10.1172/JCI26833. URL <http://dx.doi.org/10.1172/JCI26833>.
- [45] Bhavana Pothuri, Mario M. Leitao, Douglas A. Levine, Agnès Viale, Adam B. Olshen, Crispinita Arroyo, Faina Bogomolnyi, Narciso Olvera, Oscar Lin, Robert A. Soslow, Mark E. Robson, Kenneth Offit, Richard R. Barakat, and Jeff Boyd. Genetic analysis of the early natural history of epithelial ovarian carcinoma. *PloS one*, 5(4), 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0010358. URL <http://dx.doi.org/10.1371/journal.pone.0010358>.
- [46] M. T. Rae, D. Niven, A. Ross, T. Forster, R. Lathe, H. O. Critchley, P. Ghazal, and S. G. Hillier. Steroid signalling in human ovarian surface epithelial cells: the response to interleukin-1alpha determined by microarray analysis. *The Journal of endocrinology*, 183(1):19–28, October 2004. ISSN 0022-0795. doi: 10.1677/joe.1.05754. URL <http://dx.doi.org/10.1677/joe.1.05754>.
- [47] Fabien Rey, Martin H. van Vliet, Nicola J. Armstrong, Hugo M. Horlings, Karin E. de Visser, Marlen Kok, Andrew E. Teschendorff, Stella Mook, Laura van 't Veer, Carlos Caldas, Remy J. Salmon, Marc J. van de Vijver, and Lodewyk F. Wessels. A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast cancer research : BCR*, 10(6):R93+, November 2008. ISSN 1465-542X. doi: 10.1186/bcr2192. URL <http://dx.doi.org/10.1186/bcr2192>.
- [48] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, February 2007. ISSN 1460-2059. doi: 10.1093/bioinformatics/btl633. URL <http://dx.doi.org/10.1093/bioinformatics/btl633>.
- [49] M. H. Serres and M. Riley. MultiFun, a multifunctional classification scheme for escherichia coli k-12 gene products. *Microb Comp Genomics*, 5(4):205–222, 2000. ISSN 1090-6592. URL <http://view.ncbi.nlm.nih.gov/pubmed/11471834>.
- [50] M. H. Serres, S. Goswami, and M. Riley. Genprotec: an updated and improved analysis of functions of escherichia coli k-12 proteins. *Nucleic Acids Research*, 32(Database issue):D300–2, 2004. URL <http://view.ncbi.nlm.nih.gov/pubmed/11471834>.
- [51] Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, Christine Desmedt, Denis Larsimont, Fatima Cardoso, Hans Peterse, Dimitry Nuyten, Marc Buyse, Marc J. Van de Vijver, Jonas Bergh, Martine Piccart, and Mauro Delorenzi. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4):262–272, February 2006. ISSN 1460-2105. doi: 10.1093/jnci/djj052. URL <http://dx.doi.org/10.1093/jnci/djj052>.
- [52] Corey Speers, Anna Tsimelzon, Krystal Sexton, Ashley M. Herrick, Carolina Gutierrez, Aedin Culhane, John Quackenbush, Susan Hilsenbeck, Jenny Chang, and Powel Brown. *Clinical Cancer Research*, (20):6327–6340, October . ISSN 1557-3265. doi: 10.1158/1078-0432.

CCR-09-1107.

- [53] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506580102. URL <http://dx.doi.org/10.1073/pnas.0506580102>.
- [54] Roman L. Tatusov, Natalie D. Fedorova, John D. Jackson, Aviva R. Jacobs, Boris Kiryutin, Eugene V. Koonin, Dmitri M. Krylov, Raja Mazumder, Sergei L. Mekhedov, Anastasia N. Nikolskaya, B. Sridhar Rao, Sergei Smirnov, Alexander V. Sverdlov, Sona Vasudevan, Yuri I. Wolf, Jodie J. Yin, and Darren A. Natale. The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, 4(1):41+, September 2003. ISSN 1471-2105. doi: 10.1186/1471-2105-4-41. URL <http://dx.doi.org/10.1186/1471-2105-4-41>.
- [55] The.gene_ontology_consortium. Creating the gene ontology resource: design and implementation. *Genome Res.*, 11(8):1425–1433, August 2001. ISSN 1088-9051. doi: 10.1101/gr.180801. URL <http://dx.doi.org/10.1101/gr.180801>.
- [56] A. R. Thorner, K. A. Hoadley, J. S. Parker, S. Winkler, R. C. Millikan, and C. M. Perou. In vitro and in vivo analysis of B-Myb in basal-like breast cancer. *Oncogene*, 28(5):742–751, February 2009. ISSN 1476-5594. doi: 10.1038/onc.2008.430. URL <http://dx.doi.org/10.1038/onc.2008.430>.
- [57] Margaret E. Tome, David B. Johnson, Lisa M. Rimsza, Robin A. Roberts, Thomas M. Grogan, Thomas P. Miller, Larry W. Oberley, and Margaret M. Briehl. A redox signature score identifies diffuse large b-cell lymphoma patients with a poor prognosis. *Blood*, 106(10):3594–3601, November 2005. ISSN 0006-4971. doi: 10.1182/blood-2005-02-0487. URL <http://dx.doi.org/10.1182/blood-2005-02-0487>.
- [58] Melissa A. Troester, Jason I. Herschkowitz, Daniel S. Oh, Xiaping He, Katherine A. Hoadley, Claire S. Barbier, and Charles M. Perou. Gene expression patterns associated with p53 status in breast cancer. *BMC cancer*, 6:276+, December 2006. ISSN 1471-2407. doi: 10.1186/1471-2407-6-276. URL <http://dx.doi.org/10.1186/1471-2407-6-276>.
- [59] Gert G. Van den Eynden, Steven J. Van Laere, Ilse Van der Auwera, Leen Gilles, J. Lance Burn, Cecile Colpaert, Peter van Dam, Eric A. Van Marck, Luc Y. Dirix, and Peter B. Vermeulen. Differential expression of hypoxia and (lymph)angiogenesis-related genes at different metastatic sites in breast cancer. *Clinical & experimental metastasis*, 24(1):13–23, March 2007. ISSN 0262-0898. doi: 10.1007/s10585-006-9049-3. URL <http://dx.doi.org/10.1007/s10585-006-9049-3>.
- [60] Matthew G. Vander Heiden, Lewis C. Cantley, and Craig B. Thompson. Understanding the warburg effect: The metabolic requirements of cell proliferation. *Science*, 324(5930):1029–1033, May 2009. ISSN 1095-9203. doi: 10.1126/science.1160809. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2849637/>.
- [61] Christopher M. Williams, Angelo G. Scibetta, J. Karsten Friedrich, Monica Canosa, Chiara Berlato, Charlotte H. Moss, and Helen C. Hurst. AP-2gamma promotes proliferation in breast tumour cells by direct repression of the CDKN1A gene. *The EMBO journal*, 28(22):3591–3601, November 2009. ISSN 1460-2075. doi: 10.1038/emboj.2009.290. URL <http://dx.doi.org/10.1038/emboj.2009.290>.
- [62] Pratyaksha Wirapati, Christos Sotiriou, Susanne Kunkel, Pierre Farmer, Sylvain Pradervand, Benjamin H. Kains, Christine Desmedt, Michail Ignatiadis, Thierry Sengstag, Frederic Schutz, Darlene Goldstein, Martine Piccart, and Mauro Delorenzi. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research*, 10(4):R65+, July 2008. ISSN 1465-5411. doi: 10.1186/bcr2124. URL <http://dx.doi.org/10.1186/bcr2124>.
- [63] Charles E. Wood, Jay R. Kaplan, M. Babette Fontenot, J. Koudy Williams, and J. Mark Cline. Endometrial profile of tamoxifen and low-dose estradiol combination therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 16(3):946–956, February 2010. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-09-1541. URL <http://dx.doi.org/10.1158/1078-0432.CCR-09-1541>.
- [64] Matthew D. Young, Matthew J. Wakefield, Gordon K. Smyth, and Alicia Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology*, 11(2):R14+, February 2010. ISSN 1465-6914. doi: 10.1186/gb-2010-11-2-r14. URL <http://dx.doi.org/10.1186/gb-2010-11-2-r14>.
- [65] Xu-Yun Y. Zhao, Ke-Wen W. Zhao, Yi Jiang, Meng Zhao, and Guo-Qiang Q. Chen. Synergistic induction of galectin-1 by CCAAT/enhancer binding protein alpha and hypoxia-inducible factor 1alpha and its role in differentiation of acute myeloid leukemic cells. *The Journal of biological chemistry*, 286(42):36808–36819, October 2011. ISSN 1083-351X. doi: 10.1074/jbc.M111.247262. URL <http://dx.doi.org/10.1074/jbc.M111.247262>.
- [66] Yuan Zhao, Tianhua Zhou, Aiqing Li, Haomi Yao, Fei He, Liangjing Wang, and Jianmin Si. A potential role of collagens expression in distinguishing between premalignant and malignant lesions in stomach. *Anatomical record (Hoboken, N.J. : 2007)*, 292(5):692–700, May 2009. ISSN 1932-8494. doi: 10.1002/ar.20874. URL <http://dx.doi.org/10.1002/ar.20874>.
- [67] Heddy Zola, Bernadette Swart, Alison Banham, Simon Barry, Alice Beare, Armand Bensussan, Laurence Boumsell, Chris D Buckley, Hans-Jörg J. Bühring, Georgina Clark, Pablo Engel, David Fox, Bo-Quan Q. Jin, Peter J. Macardle, Fabio Malavasi, David Mason, Hannes Stockinger, and Xifeng Yang. CD molecules 2006–human cell differentiation molecules. *Journal of immunological methods*, 319(1-2):1–5, January 2007. ISSN 0022-1759. doi: 10.1016/j.jim.2006.11.001. URL <http://dx.doi.org/10.1016/j.jim.2006.11.001>.